



A Study on Hierarchical Clustering Algorithms

¹D.Saravanan, ²Dr. Dennis Joseph

Faculty of Operations & IT, ICFAI Business School (IBS), Hyderabad,
The ICFAI Foundation for Higher Education (IFHE)
(Deemed to be University u/s 3 of the UGC Act 1956)
Hyderabad, India.

Abstract: Video Content is always huge by itself with abundant information. Extracting explicit semantic information has been extensively investigated such as object detection, structure analysis and event detection. However, little work has been devoted on the problem of discovering global or inexplicit information from the huge video stream. The video is a particular media embedding visual, motion, audio and textual information. The indexing process must be automated in order to build a dictionary of images region. This process is carried out in various steps. One such important step is Clustering, which is data mining is the process of discovering groups in a dataset. In this paper, we attempt to give a comparative study of existing algorithms suitable for video data mining.

Keywords: Hierarchical. Clusterina. Video data minina. Imaae Processina. Chameleon Alaorithm.

1. Introduction

Techniques to understand digital video content becomes a hot research topic in recent years. Their main focus is on structure analysis [1], event detection [2], highlight summary [3], et al. As video contents normally take a long period of time to play and occupy a large volume of bytes to store, mining useful statistical information from the whole video content to help users have a better understanding of it is a relatively difficult task and has not been fully explored yet. Compared with object-, structure- or event-based video semantic information, statistical mining information is hard to be acquired by human labeling. Data mining is the process of finding useful patterns or extracting previously unknown knowledge from a massive set of data [4]. Different from textual information which has been studied for a long time, video contents have special characteristics: they are continuous sequences with temporal relations among them, and each video segment normally contains abundant information itself. Due to the semantic gap between human perception and computer-centered low-level features, many video understanding techniques such as shot boundary detection and event extraction remain open problems. Thus investigation on video mining is at its early stage as mining solutions normally need the extracted semantic information. In fact, the current status of video mining is still at the pre-processing stage, such as video clustering [5]. The main motivation of video mining is to find undiscovered knowledge from the stream based on visual and audio cues. The knowledge may typically include structure information within a video clip or association information among various clips, as well as trend information based on the analysis for a massive size of video set[7][8][9]. Clustering is an unsupervised machine learning process that creates clusters such that data points inside a cluster are close to each other, and also far apart from data points in other clusters. Clustering in data mining [SADC93, CHY96] is a discovery process that groups a set of data such that the intracluster similarity is maximized and the intercluster similarity is minimized [JD88, KR90, PAS96, CHY96]. Clustering is used in many applications, as a stand-alone tool to get insight into data distribution and as a preprocessing step for other algorithms. Specifically it is used in Pattern Recognition, Spatial Data Analysis, Image Processing, Economic Science, document classification etc. [10]

II. Categories of clustering algorithms

There are four main categories of clustering algorithms: partitioning, density-based, grid-based, and hierarchical.

A. Clustering Approaches

B. Partitioning algorithms

Construct various partitions and then evaluate them by some criterion (k-means, k-medoids)

C. Hierarchical algorithms

Create a hierarchical decomposition of the set of data (or objects) using some criterion (AGNES, DIANA)

D. Density-based

Based on connectivity and density functions – grow a cluster as long as density in the neighborhood exceeds a threshold (DBSCAN, CLIQUE)

E. Grid-based

Based on a multiple-level grid structure (i.e., quantized space) (STING, CLIQUE)

F. Model-based

A model is hypothesized for each of the clusters and the idea is to find the best fit of the data to the given model (EM)

III. Clustering algorithms suitable for Video data mining

Current Video data mining techniques emphasize on pattern discovery. We require content-adaptive pattern discovery techniques that would adapt to variations in content, and make the event-search tractable. Clustering in data mining is a discovery process that groups a set of data. Out of the four clustering approaches, Hierarchical algorithms are best suited for video data mining due to its simplicity and efficiency [11]. They produce a nested sequence of clusters, with a single all-inclusive cluster at the top and single point clusters at the bottom. Hierarchical clustering algorithms are classified as agglomerative or divisive. The agglomerative (bottom up) approach repeatedly merges two clusters, while the divisive (top down) approach repeatedly splits a cluster into two. Agglomerative clustering method does not suit video data mining application due to following reasons – they do not scale well, time complexity of at least $O(n^2)$, where n is the number of total objects, can never undo (backtrack) what was done previously and do not give rise to dead-end solutions. [12] Divisive clustering method is suitable for statistical analysis. This method cannot be used for video data mining as they can never undo. Hence a best suited clustering algorithm for video data mining would be to integrate Hierarchical clustering with another clustering technique. Many algorithms were proposed like BIRCH, CURE, CHAMELEON etc. Integrating Hierarchical and distance based clustering methods. In the following table, we bring out the comparative study of various clustering algorithms that are suitable for video data mining. For our purpose, we have considered BIRCH, CURE and CHAMELEON.

IV. Comparative study of Hierarchical clustering Algorithms

BIRCH	CURE	CHAMELEON
Uses tree structure and incrementally adjusts the quality of sub clusters	Selects well scattered points from cluster and then shrinks them towards the centre of the cluster by a specific fraction	Uses dynamic modeling
BIRCH employs Clustering Feature and Clustering Features tree	Cure employs combinations of random sampling and partitioning	CHAMELEON employs finding initial sub cluster and merging sub clusters using dynamic framework
Clusters multi-dimensional metric data points	Clusters metric space dataset	Can be applied to any data set
Method of forming cluster Employs tree structure and tunes the quality of sub clusters. Method of forming Cluster Employs tree structure and tunes the quality of sub clusters.	Method of forming cluster Draws well dispersed objects from the cluster and then moves them towards the cluster center by a certain fraction Method of forming Cluster Draws well dispersed objects from the cluster and then moves them towards the cluster center by a certain fraction	Method of forming cluster Uses dynamic modeling Method of forming Cluster Uses dynamic modeling
Algorithm Incrementally construct a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data) Phase 2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree	Algorithm Draw random sample s . Partition sample to p partitions with size s/p Partially cluster partitions into s/pq clusters Eliminate outliers By random sampling If a cluster grows too slow, eliminate it. Cluster partial clusters. Label data in disk	Algorithm 1. Use a graph partitioning algorithm: cluster objects into a large number of relatively small sub-clusters 2. Use an agglomerative hierarchical clustering algorithm: find the genuine clusters by repeatedly combining these sub-clusters
Improvements required: 1. Way of increasing the threshold dynamically 2. The dynamic adjustment of outlier criteria 3. more accurate quality measurements 4. Handles only numeric data and sensitive to the order of the data record 5. A new architecture to propose a parallel execution as well as interactive learning's.	Improvements required: Cure performs multiple mistakes for which remedy is sought. The following were some of the mistakes: 1. CURE failed to find the right clusters for a given data set 2. It produces a dendrogram of possible clustering solution at different levels of granularity 3. It merges together sub-clusters that belong to two different genuine clusters.	Improvements required: 1. It should support static model also. 2. Processing cost should be reduced 3. It should be applied to different applications domain and tested.

V. Conclusion

Here, we have made an attempt to compare the existing clustering algorithms that are suitable for video data mining. This study brings out the characteristics of BIRCH, CURE and CHAMELEON and focuses on the improvements needed in each algorithm. This would definitely aid in the development of new algorithms suitable for video data mining

VI REFERENCES

- [1] Ingham Oh, Babitha Bandi 'Multimedia Data Mining Framework for Raw Video sequence 'in proc MDM/KDD 2002:International workshop on multimedia data mining (with ACM SIGKDD 2002)
- [2] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image Indexing Using Color Correlograms. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97), pages 762–768, San Juan, Puerto Rico, June 1997
- [3] D.Saravanan, "Video data image retrieval using – BRICH", World journal of Engineering, Vol.14, Issuu 4, Pages 318-323, Aug 2017.
- [4] D.Saravanan" Effective Video Data Retrieval Using Image Key frame selection", Advances in Intelligent Systems and computing , Pages 145-155,jan-2017
- [5] P. Kerminen and M. Gabbouj. The Visual Goodness Evaluation of Color based Retrieval Processes. In Proc. of the EUSIPCO 2000, pages 2153– 2156, Tampere, Finland, September 5–8 2000.
- [6] O. Guldogan M. Partio, E. Guldogan and M. Gabbouj. Applying Texture and Color Features to Natural Image Retrieval. In Proc. of FINSIG'03 - 2003 Finnish Signal Processing Symposium, pages 199–203, Tampere, Finland, May 2003.
- [7] JngHwan Oh, Babitha Bandi 'Multimedia Data Mining Framework for Raw Video Sequences' In Proc. MDM/KDD 2002: International Workshop on Multimedia Data Mining(with ACM SIGKDD 2002)
- [8] Ian Davindson and S.S Ravi 'Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results '4th SIAM Data Mining Conf .2004
- [9] D.Saravanan," Effective video Content Retrieval using image attributes", EAI Endorsed Transactions on energy Web and Information Technologies, Vol5, Issues18, e8, Pages 1-5. June 2018.
- [10] D.Saravanan," Efficient Video indexing and retrieval using hierarchical clustering techniques", Advances in Intelligence systems and computing, Volume 712, Pages 1-8, ISBN:978-981-10-8227,6, Nov-2018.
- [11] L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 1990.
- [12] T. Zhang, R.Ramakrishnan, and M.Livny, "BRICH: An Efficient data clustering method for very large databases ," in Proceedings of 1996 ACM SIGMOD International Conference on Management of Data (SIGMOD '96) (Monteral , Canada) PP 103-114 , June 1996.
- [13] S.Guha, R.Rastogi, and .Shim, "An efficient clustering algorithm for large database," in proceedings of 1998 ACM_SIGMOD International conference on Management of Data (SIGMOD'98) (Seattle, WA), pp 73-84, June 1998.
- [14] G.Karypis, E.H.Han and V.Kumar "CHAMELEON: A hierarchical clustering algorithm using dynamic modeling," Computer, Vol 32pp 68-75, 1999.
- [15] D.Saravanan, "Clustering of video in formations using BRICH Methodology" Pak journal of biotechnology, Vol 14(Special Issue-II-2017), Pages 377-380, Dec- 2017
- [16] D.Saravanan, "Improved image retrieval using image noise removal technique" Pak journal of biotechnology, Vol 14(Special Issue-II-2017), Pages 360-362, Dec- 2017.
- [17] D.Saravanan, "Clustering the irregularity events in intelligence surrounding systems" Int. Journal of pure and applied mathematics, Vol. 119, No.12(2018), Pages 15025-15035, May-2018 (Special Issues) , ISSN:1311-8080.
- [18] M.Stonebaker, R.Agrawa, U.Dayal, E.J Neuhold, and A.Reuter, DBMS research at a crssroads: The Vienna update. In proc .of the 19th VLDB Conference, pages 688-692,Dublin, Ireland,1993.